A Local Rigid Body Framework for Global Optimization of Biomolecules

Halim Kusumaatmaja,* Chris S. Whittleston, and David J. Wales*

University Chemical Laboratories, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U. K.

E-mail: hk286@cam.ac.uk; djw34@cam.ac.uk

Abstract

We present a local rigid body framework for simulations of biomolecules. In this framework, arbritrary sets of atoms may be treated as rigid bodies. Such groupings reduce the number of degrees of freedom, which can result in a significant reduction of computational time. As benchmarks, we consider global optimization for the tryptophan zipper (trpzip 1, 1LE0; using the CHARMM force field) and chignolin (1UAO; using the AMBER force field). We use a basin-hopping algorithm to find the global minima and compute the mean first encounter time from random starting configurations with and without the local rigid body framework. Minimal groupings are used, where only peptide bonds, termini and side chain rings are considered rigid. Finding the global minimum is 4.2 and 2.5 times faster, respectively, for trpzip 1 and chignolin, within the local rigid body framework. We further compare $O(10^5)$ low-lying local minima to the fully relaxed unconstrained representation for trpzip 1 at different levels of rigidification. The resulting Pearson correlation coefficients, and thus the apparent intrinsic rigidity of the various groups, appear in the following order: side chain rings $>$ termini $>$ trigonal planar centres $\geq$ peptide bonds $\gg$ side chains. This approach is likely to be even more beneficial for structure prediction in larger biomolecules.

*To whom correspondence should be addressed

1 Introduction

Computer simulations have become an increasingly popular research tool in the quest to understand biomolecules and soft matter. These systems often involve physical and chemical processes that span multiple length and time scales, and exhibit very rich and interesting behaviour. Unfortunately, the very same issue of multiple length and time scales also limits what we can simulate in practice. To overcome this issue, various strategies have been developed. Advanced sampling techniques, such as replica exchange,[1] Wang-Landau,[2] umbrella sampling,[3] and metadynamics,[4] make it possible to obtain thermodynamic information, which would otherwise be available only from much longer simulations. Coarse-graining[5–7] allows us to achieve longer length and time scales by integrating out less relevant microscopic details. Significant effort has also been directed towards multiscale modelling, where various parts of the computational domain are represented at different resolution,[8–10] or where large-scale cooperative motion is first explored, followed by relaxation at an atomistic level.[11,12]

We describe here a local rigid body framework, where arbitrary sets of atoms may be grouped as rigid bodies. This procedure reduces the number of degrees of freedom during simulations or geometry optimization, thus providing a significant gain in computational efficiency. However, in contrast to coarse-graining, the full atomistic representation is still accounted for in terms of interactions between the locally rigid bodies.

The above framework can be employed for a wide variety of simulations, and here it will be discussed and implemented in the context of basin-hopping global optimization.[13–16] Basin-hopping has been succesfully applied for locating the global minima in a wide range of problems, including atomic and molecular clusters, glass-formers, and biomolecules. As our interest turns to larger systems, all-atom representations become prohibitively expensive after more than a few thousand atoms. However, since in many cases groups of atoms move very little with respect to each other (i.e. there is some intrinsic rigidity), or we are simply not interested in their displacements, local rigidification becomes an attractive approach.

Naturally the gain in computational efficiency is closely related to the reduction in the num-

ber of degrees of freedom, and hence the atoms one can group as rigid bodies. Such information may be obtained from short runs with fully atomistic simulations by performing root mean-square displacement or principal component analysis.[17,18] There are also methods, such as the pebble game,[19] where graph theoretical techniques are employed to analyze bond networks and thus rigidity in different parts of the molecules of interest. Alternatively, the local rigid body framework may be applied recursively to larger and larger domains, and by comparing the potential energy landscape to its fully atomistic counterpart, we can establish a set of rules for what we can safely group as rigid bodies without affecting the properties of interest significantly. In the present work, we focus mainly on minimal rigid body groupings, where only peptide bonds, peptide termini and side chain rings are considered rigid. We show that, even for this minimal scheme, a significant speed-up is obtained in the mean first encounter time for finding the global minimum.

This paper is organized as follows. In the next section, we detail our simulation setup and the implementation of the local rigid body framework. This account is then followed by the presentation of benchmark results for the tryptophan zipper (trpzip 1)[20] and chignolin.[21] We use a CHARMM force field[22,23] for the former peptide and an AMBER force field[24,25] for the latter to demonstrate that this approach can easily be implemented for different potentials. To justify our rigid body groupings, we present a correlation analysis between the low-lying local minima found in rigidified and unconstrained minimization at different levels of rigidification for trpzip 1. Finally, we summarize the most important findings and discuss avenues for future research.

2 Methodology

The scheme in Fig. 1 summarizes the methodology, which is based upon basin-hopping global optimization.[13–16] In this approach, after a trial move is proposed, it is followed by an energy minimization; the move is then accepted or rejected based upon the change in the corresponding energy for the local minimum, V. A simple, yet quite effective, approach is to use a Metropolis acceptance criterion: a step is accepted if $V_{\mathrm{new}} < V_{\mathrm{old}}$, or if $V_{\mathrm{new}} > V_{\mathrm{old}}$ and $\exp\{(V_{\mathrm{old}} - V_{\mathrm{new}})/kT_{\mathrm{bh}}\}$

is larger than a random number drawn from the range [0,1]. Here we use $T_{\mathrm{bh}} = 750$ K, which corresponds to an energy of 1.5 kcal/mol. Since the energy is minimised after the proposed move, the geometric perturbations proposed as steps can generally be much larger than the displacements used in typical Monte Carlo sampling for thermodynamic properties. All the results presented here were obtained using the GMIN program,[26] which is available for use under the GNU public license.

2.1 Trial moves

We have used two different types of trial move: *dihedral* moves for the global optmization of trpzip 1 and *group rotations* for chignolin. Different trial moves, such as short molecular dynamics runs, random displacements, or activated events[27] either in Cartesian or internal coordinates, could also be used and combined. For the dihedral moves, a number of backbone dihedral angles are selected and twisted in the manner described in reference.[28] To account for the effect of the dihedrals' position along the chain, we choose a lower selection probability in the middle, $P_{\mathrm{min}} = 0.2$, than at the end, $P_{\mathrm{max}} = 0.4$. We limit the maximum number of dihedral angles that can be shifted to fifteen, and the amplitude of the dihedral twist is sampled randomly between 0 and $\pi/6$.

For the group rotation moves, we rotate groups of atoms as rigid bodies.[29] The axis of rotation, group members, selection probability, and maximum rotation amplitude of each group are all input parameters. Here we apply group rotation moves to perturb the termini as well as the side chains of the peptide. As a general rule, the bigger and bulkier the group, the smaller the selection probability and amplitude of rotation employed. The detailed parameters we have used are included in the supporting information, and examples will be available online from the GMIN web site.[26]

One other basin-hopping feature of GMIN that we have used in this paper is the *newrestart* procedure. Here, we reseed the configuration from a short high temperature (T_{nst}) molecular dynamics run if the energy of the system has not decreased after a specified number of steps, N_{nst}. This procedure allows the system to escape more easily from sets of minima that act as a trap. We used $N_{\mathrm{nst}} = 5000$ and $T_{\mathrm{nst}} = 1250\,\mathrm{K}$ (2.5 kcal/mol) in the present work.

2.2 Minimization step

In the local rigid body framework, there are two features that we would like to take advantage of during the minimization step. First, we want to use widely available force fields, such as CHARMM[22,23] and AMBER,[24,25] and as such, the energy and its first derivatives are calculated using the atomistic coordinates, $\{\mathbf{r}_i\}$, native to the force field. Second, we want to exploit the reduction in the number of degrees of freedom as we group sets of atoms into rigid bodies, $\{\mathbf{r}_i\}_{i\in I} \rightarrow \{\mathbf{r}_I, \mathbf{p}_I\}$. This reduction requires us to convert the atomistic coordinates of the molecule into its rigid body coordinates, and vice versa, and to project the forces on the atoms to obtain the corresponding forces and torques on the rigid bodies. In all cases, the energy minimization was caried out using a slightly modified version of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm.[30,31]

2.3 Local rigid bodies

Each rigid body has six degrees of freedom: three representing the position of the centre of geometry (translational degrees of freedom) $\{\mathbf{r}_I\}$, and three representing its orientation (rotational degrees of freedom) $\{\mathbf{p}_I\}$. One may also use the centre of mass rather than the centre of geometry. For clarity, we employ capital letters for rigid bodies, and lower case for atoms. We will also denote $\{\mathbf{r}_i^0\}_{i\in I}$ as the reference coordinates of the atoms in rigid body I relative to the centre of geometry. Unless specified otherwise, we use the atomic configuration at the global minimum for $\{\mathbf{r}_i^0\}_{i\in I}$.

We use an angle-axis framework[32,33] to represent the rotational degrees of freedom of the rigid bodies. Here, $\mathbf{p}_I = \theta_I \hat{\mathbf{p}}_I = \left(p_I^1, p_I^2, p_I^3\right)$, where $\hat{\mathbf{p}}_I$ is a unit vector defining the rotation axis and θ_I is the magnitude of the rotation about this axis. p_I^1, p_I^2, and p_I^3 are the components of the rotation vector in the x, y and z directions. Following Rodrigues' rotation formula,[34] the rotation matrix $\mathbf{R}_I$ can be expressed as

$$\mathbf{R}_I = \mathbf{I} + (1 - \cos\theta_I)\widetilde{\mathbf{p}}_I\widetilde{\mathbf{p}}_I + \sin\theta_I\widetilde{\mathbf{p}}_I, \tag{1}$$

where $\mathbf{I}$ is the identity matrix and $\widetilde{\mathbf{p}}_I$ is the skew-symmetric matrix obtained from the rotation vector $\mathbf{p}_I$:

$$\widetilde{\mathbf{p}}_I = \frac{1}{\theta_I} \begin{pmatrix} 0 & -p_I^3 & p_I^2 \\ p_I^3 & 0 & -p_I^1 \\ -p_I^2 & p_I^1 & 0 \end{pmatrix}. \tag{2}$$

Using this rotation matrix, the mapping from rigid body to atomistic coordinates can be performed using the following formula:

$$\mathbf{r}_{i\in I} = \mathbf{r}_I + \mathbf{R}_I \times \mathbf{r}^0_{i\in I}. \tag{3}$$

The reverse transformation from atomistic to rigid body coordinates may be performed in several ways. Here we use the following procedure. The centre of geometry is given by

$$\mathbf{r}_I = \sum_{i\in I}^{n_i} \mathbf{r}_i / n_I, \tag{4}$$

where n_I is the number of atoms in rigid body I. Now we need to choose at least two atoms within each rigid body to construct three orthogonal unit vectors $(\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3)$. Let us denote the positions of these atoms relative to the centre of geometry as $\mathbf{r}_1$ and $\mathbf{r}_2$ in the rotated frame, and $\mathbf{r}^0_1$ and $\mathbf{r}^0_2$ in the reference frame. The two vectors, i.e. $\mathbf{r}_1$ and $\mathbf{r}_2$ or $\mathbf{r}^0_1$ and $\mathbf{r}^0_2$, must not be parallel. This condition implies that we need at least three non-collinear atoms in each rigid body. Three orthogonal unit vectors can then be constructed as

$$\begin{aligned} \hat{\mathbf{e}}_1 &= \mathbf{r}_1/|\mathbf{r}_1|, \\ \hat{\mathbf{e}}_2 &= \mathbf{r}_1 \times \mathbf{r}_2/|\mathbf{r}_1 \times \mathbf{r}_2|, \\ \hat{\mathbf{e}}_3 &= \hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2, \end{aligned} \tag{5}$$

for the rotated frame, and similarly for $\left(\hat{\mathbf{e}}^0_1, \hat{\mathbf{e}}^0_2, \hat{\mathbf{e}}^0_3\right)$ in the reference frame. Using the orthogonal

vectors in the rotated and reference frames, we can compute the rotation matrix as

$$\mathbf{R}_I = \begin{pmatrix} \hat{e}_{1x} & \hat{e}_{2x} & \hat{e}_{3x} \\ \hat{e}_{1y} & \hat{e}_{2y} & \hat{e}_{3y} \\ \hat{e}_{1z} & \hat{e}_{2z} & \hat{e}_{3z} \end{pmatrix} \begin{pmatrix} \hat{e}^0_{1x} & \hat{e}^0_{2x} & \hat{e}^0_{3x} \\ \hat{e}^0_{1y} & \hat{e}^0_{2y} & \hat{e}^0_{3y} \\ \hat{e}^0_{1z} & \hat{e}^0_{2z} & \hat{e}^0_{3z} \end{pmatrix}^{-1} . \tag{6}$$

The conversion from rotation matrix to angle-axis coordinates is straightforward,[34] and will not be detailed here.

In addition to transformation of coordinates, the local rigid body framework requires the projection of forces (first derivatives) from individual atoms onto the translational and rotational degrees of freedom of the rigid bodies. These projections are given by

$$\frac{\partial U}{\partial r_I^k} = \sum_{i \in I} \frac{\partial U}{\partial r_i^k}, \tag{7}$$

$$\frac{\partial U}{\partial p_I^k} = \sum_{i \in I} \nabla_i U \cdot \left(\mathbf{R}_I^k \mathbf{r}_i^0 \right), \tag{8}$$

where, following Eq. (3), we have used

$$\frac{\partial \mathbf{r}_i}{\partial p_I^k} = \mathbf{R}_I^k \mathbf{r}_i^0. \tag{9}$$

2.4 The tryptophan zipper (trpzip 1)

The β-hairpin-forming peptide tryptophan zipper (trpzip 1, PDB: 1LE0[20]) has 12 amino acid residues; the global minimum energy configuration is illustrated in Fig. 2(a). Its sequence is SER - TRP - THR - TRP - GLU - GLY - ASN - LYS - TRP - THR - TRP - LYS. We used the CHARMM19 force field and the EEF1 implicit solvent model[35] for our global optimization analysis of this peptide. We symmetrized the CHARMM force field so that the potential is invariant under feasible permutations of identical atoms.[36] The first and last residues were patched using the *NTER* and *CT2* options in CHARMM, respectively. This procedure is needed to adjust the connectivity of the

terminal residues in the simulations, since they have one neighbor fewer than those in the centre. For the EEF1 implicit solvent parameters, we set the temperature in the model to 298.15 K. The electrostatic energy is taken to be inversely proportional to the square of the atom-atom pair distance (group RDIE), and the following cutoff values were used: ctonnb 7.0, ctofnb 9.0, and cutnb 10.0.

There are in total 147 atoms (441 degrees of freedom) in the system. By grouping the termini, peptide bonds, and tryptophan side chain rings as shown in Fig. 2(c), we can reduce the number of degrees of freedom by 40% to 264 (17 rigid bodies and 54 ungrouped atoms).

2.5 Chignolin

Chignolin[21,37] is a de novo peptide, which was designed to form a β-hairpin. It has ten amino acid residues and its sequence is GLY - TYR - ASP - PRO - GLU - THR - GLY - THR - TRP - GLY. While the reported PDB structures (PDB: 1UAO) are local minima for the AMBERff03 force field with a generalised Born implicit solvent model,[38] we note that they do not correspond to the global minimum configuration, and similarly for the AMBER99SB force field. The global minimum structure is instead the one shown in Fig. 2(b) for the AMBERff03 force field, and it is this minimum that we searched for in our global optimization runs.

We used the Onufriev-Bashford-Case (OBC) generalised Born implicit solvent model (*igb* = 2)[38] with no periodic boundary condition. No cutoff was used for the nonbonded interactions by setting *cut* = 999.99. The maximum distance between atom pairs considered for effective Born radii calculations was set at 25.0 (*rgbmax*). We have also symmetrized the AMBER topology file as described in reference,[36] so that the potential is invariant with respect to feasible permutations of identical atoms.

Out of 138 atoms (414 degrees of freedom) in chignolin, we grouped the peptide bonds, as well as the tyrosine, proline and tryptophan side chain rings in our local rigid body framework. The groups are shown in Fig. 2(c). This grouping reduces the number of degrees of freedom in the system by 22% to 324 (11 rigid bodies and 86 ungrouped atoms).

3 Mean first encounter time

To quantify the efficiency of our local rigid body framework, we compute the mean first encounter time required to find the global minimum from random starting configurations. The starting configurations were generated from high temperature molecular dynamics runs, each sampled over 100,000 steps with a time step of 2 fs at 750 K (1.5 kcal/mol). To provide a meaningful comparison, we use the same number of starting configurations for the local rigid body framework and for the unrigidified search, i.e. 100 configurations for trpzip 1 and 50 configurations for chignolin.

While we would ideally like to use a larger number of starting configurations, we are limited by the required computational time, especially for the unconstrained global optimizations. The first encounter time distribution typically has a long tail, and this is precisely the reason why the local rigid body framework was developed. Nonetheless, our experience with smaller systems suggests that $O(100)$ starting configurations provides a reasonable estimate for the mean first encounter time.

Our choice for T_{bh}, T_{nst}, and N_{nst} was guided using the following procedure for trpzip 1. Using ten starting configurations for each set, we ran 150,000 basin-hopping steps for T_{bh} = 500, 750, and 1000 K, T_{nst} = 1000, 1250, and 1500 K, and N_{nst} = 5000, 10000, and 15000. We then chose the parameters based on how many runs succeeded in finding the global minimum, and their corresponding time averages. While the sample size is clearly too small for full parameter optimization, our aim was simply to avoid particularly bad parameter values. We settled for T_{bh} = 750 K, T_{nst} = 1250 K, and N_{nst} = 5000, for both trpzip 1 and chignolin. The reported values therefore represent upper bounds for the mean first encounter time with GMIN.

We present statistics of the first encounter time for the global minimum for trpzip 1 and chignolin in Figs. 3 and 4, respectively. Panels (b) and (a) correspond to global optimization with and without the local rigid body framework, respectively. The x axis corresponds to the number of times the energy and gradient (function calls) were computed during the simulations. This measure is a more useful quantity to compare when different computer architectures and processors are used. As an example, two billion energy and gradient calculations (our longest unrigidified

calculation for trpzip 1) is equivalent to about 2000 CPU hours on an Opteron 280 processor with a 2.4GHz clock speed.

Apart from the reduction in degrees of freedom during minimization, the rigidified and unrigidified simulations were run using identical parameters, including the geometrical perturbations. The mean first encounter time for trpzip 1 is 10^8 (standard deviation $\sigma = 0.9 \times 10^8$) and 4.2×10^8 ($\sigma = 3.5 \times 10^8$) function calls with and without the local rigid body framework. The corresponding mean first encounter time for chignolin is 0.7×10^8 ($\sigma = 5.7 \times 10^7$) function calls with local rigid body and 1.8×10^8 ($\sigma = 1.4 \times 10^8$) function calls without the local rigid body framework. Thus, we obtain an overall speed-up of 4.2 and 2.5 for trpzip 1 and chignolin, respectively.

The computational gain obtained using the local rigid body framework can be broken down into three contributions.

- First, we do not have to compute pairwise interactions within a rigid body, since these terms are effectively set to zero.[39] This gain in efficiency scales as the square of the size of the rigid body if there are no cutoffs. For minimal rigid body groupings, this contribution is fairly small, but we expect it to become more important as we rigidify larger domains in bigger systems, such as secondary structures (alpha helices, beta sheets).

- Second, the number of LBFGS steps required to find a local minimum after each perturbation in geometry is reduced. Looking at our benchmark results for trpzip 1 (Fig. 3) and chignolin (Fig. 4), the improvement is about $15 - 20\%$. These results, as well as our preliminary calculations for larger systems, suggest that we should expect savings in computer time to scale with the reduction in the number of degrees of freedom.

- The third and last contribution is probably the most important one, and it comes from the reduced number of basin-hopping steps required to find the global minimum. For trpzip 1, global optimization using local rigid bodies is about four times faster, while for chignolin the gain is approximately a factor of two. By rigidifying groups of atoms, our search space is reduced. Local minima in which the relative configurations of atoms in group *I* differ

from $\{\mathbf{r}_i^0\}_{i\in I}$ are no longer accessible. We anticipate this third contribution to be the most significant (perhaps by an order of magnitude or more) as larger systems are investigated.

4 Correlation analysis for the tryptophan zipper

In the previous section we have established that the local rigid body framework provides a considerable reduction in computational time. To use this framework in practice, we need to address two additional complementary issues: first, which groups of atoms can be rigidified; and second, which reference coordinates $\{\mathbf{r}_i^0\}_{i\in I}$ should be chosen for those atoms. Here we use a correlation analysis between 150,000 low-lying energy minima obtained from rigidified and unrigidified global optimization to provide insight into these two issues.

Let us first consider the issue of local rigidity. We use a bottom-up approach where we rigidify different groups of atoms separately and in combination, and show how these groupings affect the energy correlation analysis. The correlation analysis may be performed in two different ways. **Scheme I**. We start with the unconstrained minimum configurations, rigidify the atoms, and reminimize the energy. **Scheme II**. Alternatively, we can start from the rigidified configurations, relax the local rigidifications, and reminimize the energy. Results for both schemes are presented in Fig. 5(a) and (b) when the tryptophan rings are rigidified for trpzip 1. In both cases, there is clearly a strong correlation between the rigidified and unrigidified minima. The black line corresponds to the case where the energies of rigidified and unrigidified minima are equal. To quantify the correlation, we compute the Pearson correlation coefficients[40] and the results are presented in Table 1. In fact, we have performed this analysis for different levels of local rigidification: (i) tryptophan rings, (ii) tryptophan rings and termini, (iii) tryptophan rings and peptide bonds, (iv) tryptophan rings and trigonal planar centres, (v) tryptophan rings, termini and peptide bonds, (vi) tryptophan rings, termini, trigonal planar centres and peptide bonds, and (vii) entire side chains.

Let us first concentrate on groupings (i) to (vi), where Scheme II produces high Pearson corre-

Table 1: Pearson correlation coefficients for 150,000 low-lying minima obtained in rigidified and unrigidified global optimization runs. The rigidified groups are: (i) tryptophan rings, (ii) tryptophan rings and termini, (iii) tryptophan rings and peptide bonds, (iv) tryptophan rings and trigonal planar centres, (v) tryptophan rings, termini and peptide bonds, (vi) tryptophan rings, termini, trigonal planar centres and peptide bonds, (vii) entire side chains.

	Scheme I	Scheme II
Rigid groups	Atomic to rigid	Rigid to Atomic
(i)	0.97	0.99
(ii)	0.95	0.98
(iii)	0.89	0.94
(iv)	0.91	0.97
(v)	0.92	0.96
(vi)	0.88	0.96
(vii)	0.33	0.92

lation coefficients. This result suggests that most, if not all, local minima found in rigidified global optimization are also minima in the unrigidified search. This conclusion is further supported by root mean square displacement (RMSD) analysis, which typically gives nearly vanishing RMSD values. As expected, the energy of the system also decreases systematically as we relax the local rigidifications in scheme II [see e.g. Fig. 5(b)].

Correlation scheme I provides the more stringent criterion. It gives a measure of the likelihood that a local minimum in the unconstrained search is also a minimum for constrained optimization. From Table 1, the Pearson correlation coefficients are high [though slightly lower than scheme II for groupings (i) to (vi)]. For this reason, we judge them safe for local rigidifications. This analysis justifies our choice of local rigid bodies presented in the previous section.

The last local rigidification grouping (side chains) is an interesting case, given that several coarse-grained models represent an amino acid residue by a (coarse-grained) bead, see e.g.[41] and the references therein. We find that the Pearson correlation coefficient is rather low for scheme I, where all-atom local minimum configurations are reoptimized upon local rigidification. The energy comparisons are shown explicitly in Fig. 5(c), and the low coefficient clearly reflects the broad scatter seen in the plot. Furthermore, while the resulting correlation coefficient is high for scheme II, the data points show a clear and systematic energy gap from the black line, where the

energies of rigidified and unrigidified optimizations would be equal. These two results suggest that significant rearragements of atoms can occur as we reoptimize minima either upon rigidification or relaxation of the rigid groups. We therefore conclude that this level of rigidification is too coarse for the purpose of structure prediction in small peptides.

Finally, we briefly comment on the choice of the reference coordinates $\{\mathbf{r}_i^0\}_{i\in I}$. In the analysis presented so far, we have simply used the atomic coordinates of the global minimum, but of course we are most interested in the case where this configuration is unknown. We have therefore carried out the same analysis presented in this section using several different reference minima, chosen randomly from the 150,000 we have collected. Apart from the last rigidification scheme (vii - side chains), the relative atomic configurations are, in fact, very similar from one minimum to another, which is another measure of rigidity. As such, we find very similar Pearson correlation coefficients, and the global minimum configuration survives, albeit with slightly different energies depending on the choice of $\{\mathbf{r}_i^0\}_{i\in I}$. In each case the true global minimum is located with essentially the same effciency.

5 Conclusions

In summary, we have presented and detailed the implementation of a local rigid body framework for simulations of biomolecules. Using global optimization for the peptides trpzip 1 (12 amino acid residues) and chignolin (10 amino acid residues), we have demonstrated that significant gains in computational efficiency may be obtained. The largest contribution comes from the reduction in the number of basin-hopping steps required to find the global minimum. Justification for local rigidification may be obtained using a recursive energy correlation analysis, where larger and larger domains are rigidified. We find that the correlation between the energy of the true local minima and those obtained with constraints decreases in the following order: rings $>$ termini $>$ trigonal planar centres $\geq$ peptide bonds $\gg$ side chains.

There are a number of avenues of future research available using this local rigid body frame-

work. Here we only mention two directions that we are currently pursuing. First is the application to larger biomolecules containing hundreds of residues (thousands of atoms). For these systems, all-atom global optimization becomes computationally expensive. The local rigid body framework provides a systematic way to investigate these systems, focusing on deformation modes of interest. Preliminary results for these larger systems indicate that a reduction in computational time by an order of magnitude is possible. We are also attempting to visualize the change in the energy landscape as larger and larger domains are rigidified. This analysis will not only be valuable for the local rigid body framework, but it will also provide insight into alternative coarse-graining schemes.

Acknowledgement

The authors thank Dr Dwaipayan Chakrabarti and Dr Victor Rühle for discussions on the angle-axis framework, Dr Sandeep Somani for assistance in setting up the CHARMM and AMBER input, and James Farrell for discussions on mean first encounter time analysis. This research was funded by EPSRC Programme grant EP/I001352/1 and ERC grant RG59508.

Supporting Information Available

Examples of input and output files for the global optimization calculations presented here are available as supporting information.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.

(2) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(3) Torrie, G.; Valleau, J. *J. Comput. Phys.* **1977**, *23*, 187.

(4) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562.

(5) Gohlke, H.; Thorpe, M. F. *Biophys. J.* **2006**, *91*, 2115.

(6) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073.

(7) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819.

(8) Heath, A. P.; Kavraki, L. E.; Clementi, C. *Proteins: Struc. Func. Bioinf.* **2007**, *68*, 646.

(9) Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630.

(10) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(11) Dupuis, L.; Mousseau, N. *J. Chem. Phys.* **2012**, *136*, 035101.

(12) Dupuis, L.; Mousseau, N. *J. Phys.: Conf. Ser.* **2012**, *341*, 012015.

(13) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, 2003; p. 340-346.

(14) Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6611.

(15) Wales, D. J.; Doye, J. P. K. *J. Phys. Chem. A* **1997**, *101*, 5111.

(16) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368.

(17) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164.

(18) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Chem. Phys.* **1996**, *100*, 2567.

(19) Jacobs, D. J.; Thorpe, M. F. *Phys. Rev. Lett.* **1995**, *75*, 4051.

(20) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 5578.

(21) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. *Structure* **2004**, *12*, 1507.

(22) Brooks, B. R.; III, C. L. B.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Caflisch, S. B. A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comp. Chem.* **2009**, *30*, 1545–1615.

(23) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.* **1983**, *4*, 187–217.

(24) Ponder, J. W.; Case, D. A. *Adv. Prot. Chem.* **2003**, *66*, 27–85.

(25) Case, D.; Darden, T.; Cheatham, T.; III,; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Walker, R.; Zhang, W.; Merz, K.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A.; Kolossvai, I.; Wong, K.; Paesani, F.; Vanicek, J.; Wolf, R.; Liu, J.; Wu, X.; Brozell, S.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D.; Mathews, D.; Seetin, M.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. Amber 12. University of California, San Francisco, 2012.

(26) Wales, D. J. GMIN: A program for finding global minima and calculating thermodynamic properties. `http://www-wales.ch.cam.ac.uk/GMIN`, (accessed August 1, 2012).

(27) Yun, M.-R.; Lavery, R.; Mousseau, N.; Zakrzewska, K.; Derreumaux, P. *Proteins: Struc. Func. Bioinf.* **2006**, *63*, 967.

(28) Mortensen, P. N.; Wales, D. J. *J. Chem. Phys.* **2001**, *114*, 6443.

(29) Whittleston, C. *PhD Thesis*; University of Cambridge, 2011; p. 31-34.

(30) Nocedal, J. *Math. Comput.* **1980**, *35*, 773.

(31) Liu, D.; Nocedal, J. *Math. Program.* **1989**, *45*, 503.

(32) Chakrabarti, D.; Wales, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1970.

(33) Wales, D. J. *Phil. Trans. R. Soc. A* **2005**, *363*, 357.

(34) Brannon, R. M. Rotation. `http://www.mech.utah.edu/~brannon/gobag.html`, (accessed August 1, 2012).

(35) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133.

(36) Małolepsza, E.; Strodel, B.; Khalili, M.; Trygubenko, S.; Fejer, S.; Wales, D. *J. Comp. Chem.* **2010**, *131*, 1402.

(37) Roy, S.; Goedecker, S.; Field, M. J.; Penev, E. *J. Phys. Chem. B* **1997**, *101*, 5111.

(38) Onufriev, A.; Bashford, D.; Case, D. *J. Phys. Chem. B* **2000**, *104*, 3712.

(39) This was explored with AMBER force fields by editing the NUMBER_OF_EXCLUDED_ATOMS and EXCLUDED_ATOMS_LIST flags in the topology file. We found that the computational gain was limited by the solvent model. For the Generalized Born model, the bulk of the computer time was spent in calculating the effective Born radii of the atoms.

(40) Gibbons, J. D. *Nonparametric Statistical Inference*; Marcel Dekker: New York, 1985; p. 483-488.

(41) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

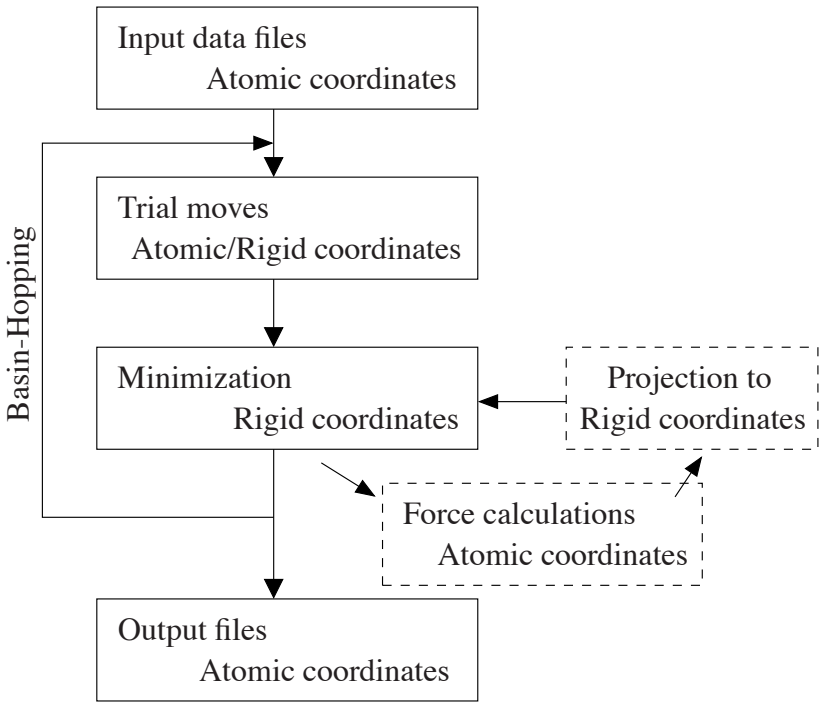

Figure 1: Schematic diagram of the basin-hopping algorithm within a local rigid body framework. Local rigidifications require mapping of atomistic positions and forces to their rigid body equivalents.

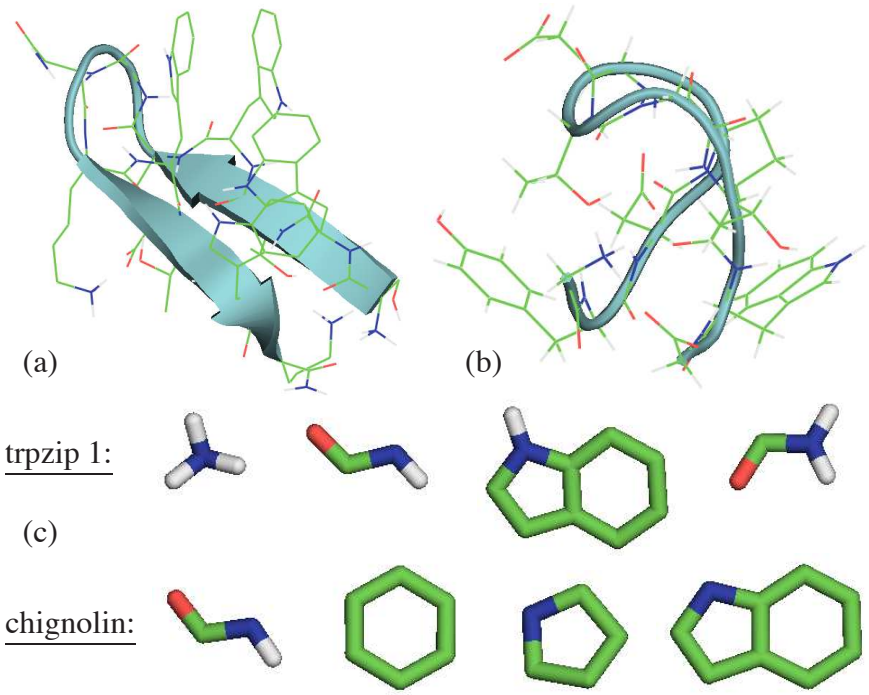

Figure 2: Global minimum configurations for (a) trpzip 1 and (b) chignolin. We also show the rigidified groups of atoms for both peptides in (c). For trpzip 1, we group the termini, peptide bonds, and tryptophan side chain rings, whereas for chignolin, we locally rigidify the peptide bonds, as well as the tyrosine, proline and tryptophan side chain rings.

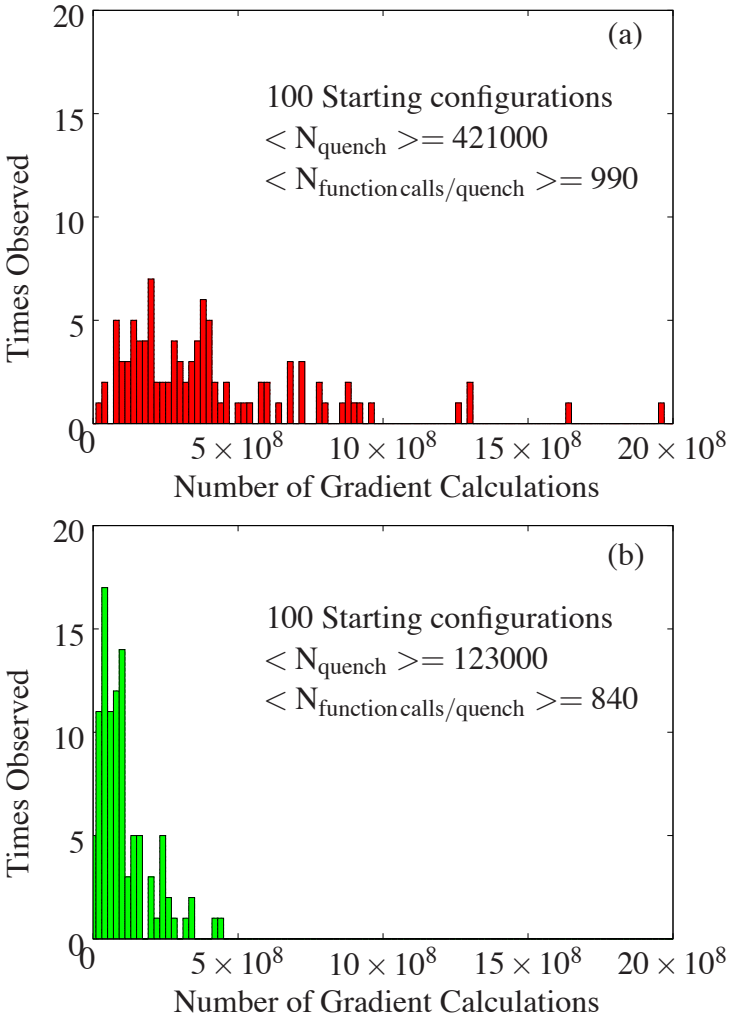

Figure 3: Distribution of first encounter times required to locate the global minimum for trpzip 1: (a) unconstrained global optimization, and (b) global optimization with local rigid bodies.

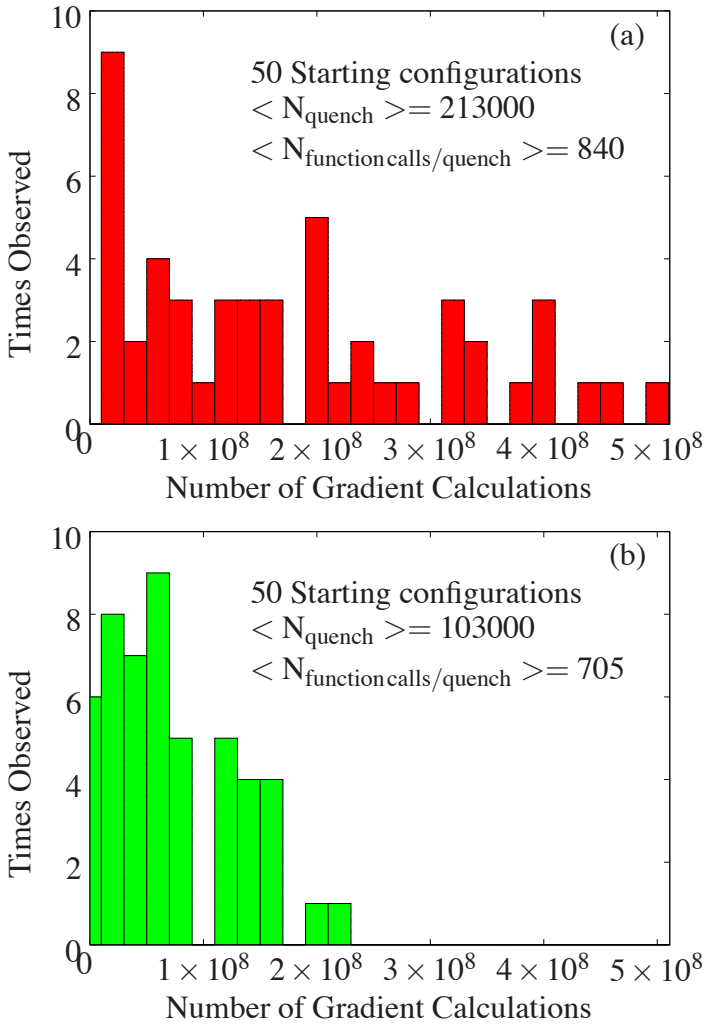

Figure 4: Distribution of first encounter times for the global minimum of chignolin: (a) unconstrained global optimization, and (b) global optimization using local rigid bodies.

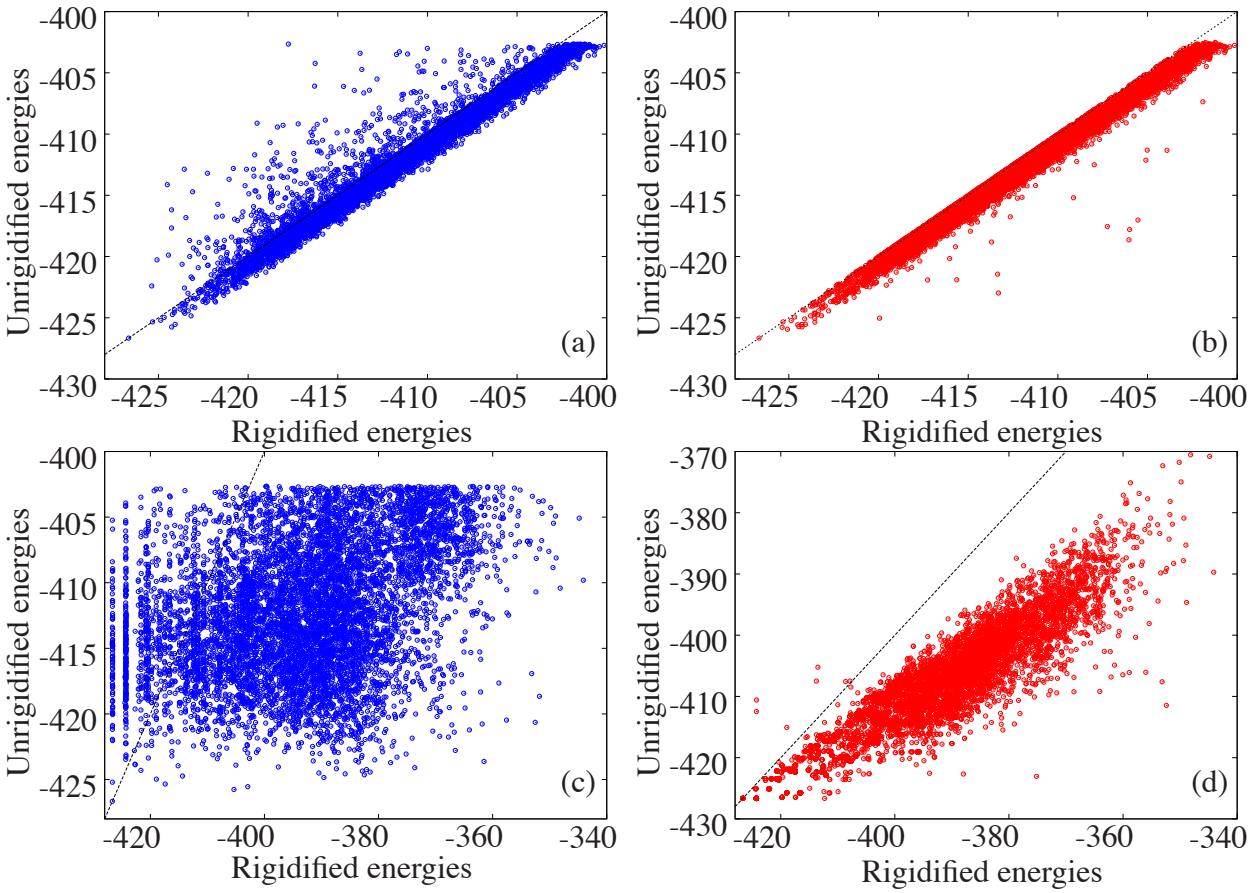

Figure 5: Comparisons between the energies of local minima obtained using rigidified and unrigidified global optimization. The left panels correspond to scheme I, and the right panels to scheme II (see text for descriptions of schemes I and II). In (a) and (b), we rigidify tryptophan rings, while in (c) and (d), we rigidify every side chain as a rigid body.